A General and Flexible Deskewing Method Based on Generalized Projection

E. DEL NINNO
E-mail: enrico.delninno@elsag.it

G. NICCHIOTTI

E. OTTAVIANI

Research & Development Department
Elsag Bailey, un'azienda di Finmeccanica S.p.A.
Via Puccini 2 16154 Genova, ITALY

A non-linear projection-based deskewing algorithm for documents and newspaper images is presented. Project profiles are among the most popular methods for skew estimation. Deskewing strategy is obtained with a three steps procedure: focus of attention selection, skew estimation and skew correction. The algorithm works on text lines and space lines projection profiles, and the dominant orientation of these lines represents the skew angle of the page.

1   Introduction

A text line is a group of characters relatively close to each other and through which a straight line can be drawn usually with horizontal or vertical, mainly for Japanese, orientation. The dominant orientation of the text lines in a document page represents the skew angle of the page. A document originally has zero skew, that is the lines are horizontally printed and often parallel to the paper edges. However when a page is manually scanned or Xeroxed, non-zero skew may be introduced. Since OCR and page layout analysis usually work on an input image with zero skew, it is important to perform skew estimation and correction before these steps. The main strategies for skew estimation techniques are:

- projection profiles [1-3]
- baseline fitting using Hough Transform [4-6]
- nearest neighbour clustering [7]

Projection profiles are among the most popular methods for skew estimation. Such strategy has proven its effectiveness in commercial systems dealing with forms. Its major shortcoming is represented by a loss of robustness if such algorithm process pages where no straight lines are present (for example newspaper pages or letters).

Hence projection profiles methods suffers for lack of generality. On the other hand Hough transform baseline fitting strategies show a very high computational burden and are much slower than profile methods though more robust. Nearest neighbour based algorithms represent a good trade-off among the two above mentioned techniques when the application does not require a good accuracy in the skew estimation. Indeed they are very often used for an initial skew angle estimate.

In this paper we will describe a new skew estimation technique belonging to the first of the three above mentioned categories. Our algorithm shows two main differences with respect to the techniques presented in literature. We use Generalised Projections (GP) to overcome the major drawbacks of projection profile methods, making more flexible and general our approach and, moreover, we may use the white pixel projection (instead of black) for newspaper page deskewing.

2 Generalised Projections

The Generalised Projections represent an extension of the well-known Projection Profiles (PP). For a binary (black and white) image $I(i, j)$ (where index i runs along the y-axis) the projection profile $P(i)$ of the row i is:

$$P(i) = \sum_j I(i, j)$$

If the value of black pixels is 1 the rows with highest PP are very likely to contain text. Anyway only if the page is already deskewed the peaks of text lines are sharp enough. The projection profiles based deskewing methods try to estimate the skew angle θ by maximising the performance index:

$$E(\vartheta) = \sum_i P_\vartheta^2(i)$$

where the PP is evaluated along straight lines at θ degrees.

Indeed such method proves to be effective and accurate only if in the text image straight lines are present. In this case the peaks of such lines are very sharp when the image is deskewed and the maximum of $E(\theta)$ is easily detectable. Anyway when no straight lines are present on the image and this is often the case for newspaper and letter pages, the maximisation of $E(\theta)$ is much more cumbersome.

The GPs extend the concept of the projection profiles introducing a "measuring function" F that allows to include into the algorithm some a-priori knowledge of the problem. The GP $G(i)$ of a row of $I(i,j)$ can be described as:

$$G(i) = \sum_j F(I(i, j))$$

As we will see the measuring function provides the algorithm with much more flexibility and accuracy.

3 Deskewing Strategy

Our deskewing algorithm consists of three main steps:

- Focus of attention selection
- Skew estimation
- Skew correction

Among these the crucial phase is of course the second. The GP algorithm is used both in the first and second step, while the third one is a simple document rotation.

3.1 Focus of Attention Selection

In its initial phase the algorithm aims to detect the region of the image more suitable for skew estimation (i.e. a focus of attention).

It makes no sense to use uniform white regions to estimate skew. Such regions give no useful contribute to the correct estimation indeed they introduce noise during the maximisation process flattening the peak relative to the correct angle. Moreover we want to understand if in the image are or are not present straight lines in order to use different "measuring functions" for the generalised projection.

To select the focus of attention with compute the GP of the rows using as measuring function F:

$$F(I(i,j)) = \begin{cases} F(I(i,j-1))+1 & \textit{if pixel is black} \\ 0 & \textit{otherwise} \end{cases}$$

This way the contribute of continuous black lines is quadratic with respect to their length, and also the peaks of text lines are sharper.

The most relevant local maxima of such generalised projections are then considered for further evaluation. All the peaks higher than a given fraction of the absolute maximum of the GP are then selected. Anyway if the maxima counter is greater than a threshold T, and this is true when we deal with letter or newspaper pages, then only the highest T peaks are considered.

Around each of such maxima R_i a region Reg_i is selected by considering all the rows lying in the interval $[R_i - \Delta_i, R_i + \Delta_i]$, where Δ_i is an adaptive length depending on the shape of the peak $GP(R_i)$.

The focus of attention area FOA is then represented by the union of such regions:

$$FOA = \bigcup_i \text{Re} \, g_i$$

We can also try to estimate if in the page are present straight lines by examining the global maxima M_g of GP. Using F as measuring function the upper bound for GP is

$$Max = \sum_{i=1}^{L} i = L(L+1)/2$$

where L is the length of the rows. Hence if $M_g > 0.3 * Max$, we guess that in the image at least one black row is present and so we use this information in the next steps.

During this phase we must pay attention in excluding the contribution of image upper and lower black borders from GP. Such borders may arise during scanning and may cause errors because their skew is not related with the true skew of text lines. This can be easily accomplished looking for the maxima R_i only between the first and last local minima.

3.2   Skew Estimation

The basic idea consists of using continuous straight line to evaluate the skew angle. Such lines are the black lines if they are present in the image, or the white lines of the baseline skip between two adjacent text lines in the case of letters or newspaper pages. In either case during the FOA selection we extracted the correct region as we detected horizontal lines if they are or text regions. If we want to use black lines the measuring function is the same F we presented above, in the case of white lines the measuring function F is:

$$F(I(i,j)) = \begin{cases} F(I(i,j-1))+1 & if \quad pixel \quad is \quad white \\ 0 & otherwise \end{cases}$$

If the GP has to be calculated along the direction θ for a line passing through the point (i, j), the previous equation can be easily extended into

$$F(I(i+\tan \vartheta * j, j)) = \begin{cases} F(I(i+\tan \vartheta(j-1), j-1))+1 & if \quad pixel \quad is \quad white \\ 0 & otherwise \end{cases}$$

This way we compute in the FOA region the GP varying the θ parameter. For each θ we can compute a performance index $V(\theta)$ given by:

$$V(\vartheta) = \sum_{FOA} GP(\vartheta)$$

The higher $V(\theta)$ the more likely θ is the correct skew angle. The search for the maximum of $V(\theta)$ is performed with a multiscale strategy. At the beginning $V(\theta)$ is maximised over a roughly undersampled version of the space of the allowed skew angle, then we refine such estimate around the initial guess.

3.3   Skew Correction

Once the skew angle θ has been estimated, the skew correction is performed with a two pass 1D transform in order to minimise the computational burden. The 1D decomposition of the rotation is represented by:

$$\begin{pmatrix} \cos \vartheta & \sin \vartheta \\ -\sin \vartheta & \cos \vartheta \end{pmatrix} \equiv \begin{pmatrix} 1 & 0 \\ -\tan \vartheta & \sec \vartheta \end{pmatrix} \begin{pmatrix} \cos \vartheta & \sin \vartheta \\ 0 & 1 \end{pmatrix}$$

Such a separable transform is computationally less expensive that a standard 2D rotation, giving an identical result on ideal images. In real cases there is a negligible degradation due to error propagation during the two passes.

4 Results

The algorithm has been tested over a large sample of document images (about one hundred). The test set contained a wide variety of documents: forms, faxes, invoices, letters, newspaper pages. The allowed skew angles are in the range (-20°, 20°). Two kinds of test have been performed.

In the first the performances of the algorithm have been evaluated in a qualitative way because we have no information about the real skew of the document images. In this case our deskewing method performs correctly for 111 of the 114 input documents, thus with a percentage of errors below 3%. All the errors occurred onto images where a large photograph was completely blackened, due to the binarisation effect. This introduces an ambiguity area for the deskewing algorithm, as, in a black area, a continuous black line can be drawn in several directions.

In the second test a document image has been numerically skewed and we can compare the results of the algorithm with respect to the actual skew.

Fig.1 - *The two images used for quantitative tests*

The results are presented in the following tables: the first one refers to a form page, the second one to a newspaper page (see Fig.1). All the data are in degrees.

Rot	0	1	2	3	4	5	6	7	8	9	10	12
Est	0	0.99	2.03	3.03	4.16	5.07	6.19	7.20	8.05	9.16	10.3	12

Average error for the form: 0.14

Rot	0	1	2	3	4	5	6	7	8	9	10	12
Est	0	0.86	2.06	2.93	3.96	5.09	5.95	6.99	8.03	9.03	10.1	12

Average error for the form: 0.06

The results demonstrate an average accuracy of the algorithm below 0.2 degrees both for forms and newspaper pages. Such accuracy is comparable with the intrinsic lower bound of the algorithm due to the pixelisation $2/L$ (rads) where L is the width in pixels of the image (order of magnitude typically about 1000).
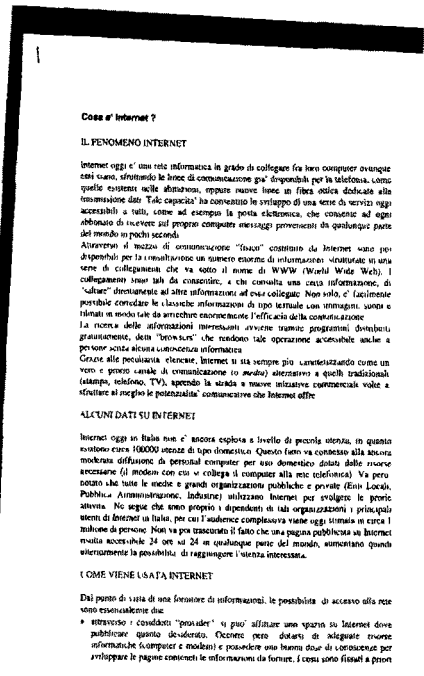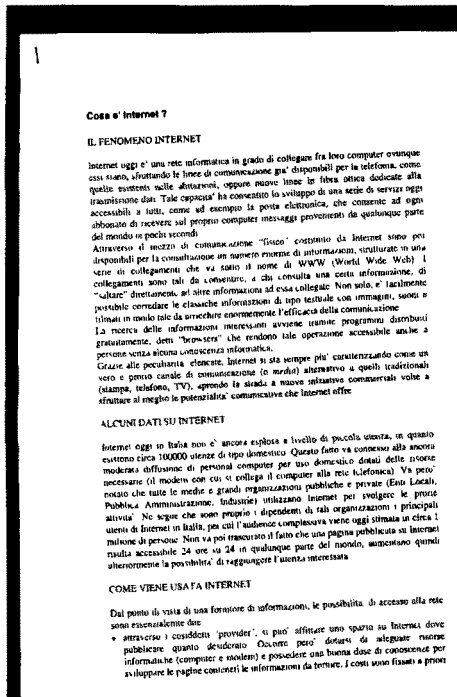

Fig.2 - A critical case

5   Further Work and Conclusions

We have presented a new eskewing method based on generalised projections. The algorithm seems to possess a wider field of application with respect to standard projection profile methods, while preserving a low computational burden if compared to Hough transform methods. The novelty of the approach is emphasised by the use of the white lines between textlines instead of the black lines as usually done. This allows

to cope with situation as the one presented below where the paper has be scanned correctly but the Xeroxing introduced a skew angle of about 2 degrees in the textlines.

The future work will be aimed to improve the performances of the algorithm with respect to pages where black areas are present due to binarisation effects on photographs. The basic idea is to exploit even the edge information of the image thus using the image gradient data. The first results of such approach seem promising allowing to overcome the ambiguity problems arising with blackened photographs.

6   References

[1] T.Akiyama and N.Hagita *"Automated entry system for printed documents"* Pattern Recognition, 23(11) ,1990 pp 1141-1154

[2] T.Pavlidis and J.Zhou, *"Page Segmentation by White Streams"* Proc 1ˢᵗ ICDAR 1991 pp 945-953

[3] H.S.Baird *"The skew angle of printed documents"* Proc. Conf. Phot. Sc. and Engin. SPIE 1987 pp 14-21

[4] S.C.Hinds et al. *"A document skew detection method using run length encoding and the Hough transform"* Proc 10ᵗʰ ICPR Los Alamitos 1990, pp 464-468

[5] Y.Nakano et al. *"An algorithm for skew normalisation of document images"* Proc. 10ᵗʰ ICPR Los Alamitos 1990, pp 8-13.

[6] B.Yu and A.K.Jain *"A robust and fast skew detection algorithm for generic documents"* Patt. Rec. 29(10) 1996 pp 1599-1629.

[7] L.O'Gorman *"The document spectrum for structural page layout analysis"* PAMI 15(11) 1993 pp 1162-1173.